2025年度湖北省科学技术奖公示表（自然科学）

项目名称、提名者及提名意见、项目简介、代表性论文专著目录、主要完成人（完成单位）

|  |  |
| --- | --- |
| 项目名称 | 深度学习模型的对抗攻击方法研究 |
| 提名单位 | 华中科技大学 |
| 提名意见 | （不超过600字，根据项目创造性特点，科学技术水平和应用情况并参照相应奖类条件写明提名理由和结论性意见，并填写提名意见和提名等级。）  项目针对深度学习模型部署推广所面临的安全可靠性问题，在国家自然科学基金等项目的支持下，在深度学习模型对样本生成的科学问题上进行深入探索，揭示了计算机视觉、自然语言处理领域的对抗样本生成机理，为发展安全鲁棒的深度学习模型及其可信落地奠定了坚实的基础。主要贡献包括：在计算机视觉领域，提出了放缩不变的对抗攻击方法和基于方差缩减的对抗攻击迁移方法，显著提高了对抗迁移攻击的成功率；在自然语言处理领域，探索了因文本与图像的不同所带来的对抗样本概念差别，定义了基于单词离散空间和考虑语义相似性的文本对抗样本的概念，建立了基于同义词替换的文本分类深度模型的对抗攻击框架并设计了基于单词显著性的黑盒对抗攻击算法。上述研究成果获得了国际学术界的广泛认可，并且被广泛引用，5篇代表作的谷歌学术引用2200余次、WOS核心合集他引1000余次、CNKI总他引1043次、单篇最高引用848次；所提出的对抗攻击方法已成为深度学习模型鲁棒性研究的重要测试基准。项目成果被ARES、CANARY、OpenAttack等多个对抗攻防鲁棒性评测平台整合；作为核心成员，参与IEEE关于模型鲁棒性测试与评估标准的制定。第一完成人多次受邀做相关大会报告。  提名该项目为2023年度湖北省自然科学奖 一 等奖 |
| 项目简介 | 本项目属于计算机科学技术领域。  深度学习模型因其极高的性能已获得广泛的应用。然而，若在原始样本数据中添加人类难以察觉的某种扰动，便能欺骗深度学习模型使之输出错误的结果。深度学习模型的安全性问题倍受业界关注，如何有效生成对抗样本以测试模型鲁棒性，从而进一步提升其安全、可信性，已成为当今业界的焦点。在国家自然科学基金等多项重要课题的支持下，本团队经过多年的积累、研究与探索，在图像识别与文本分类等深度学习模型核心应用场景中的对抗样本生成方面取得了多项突破，解决了深度学习模型对抗攻击中的若干关键科学问题，建立了深度学习对抗攻击技术体系与安全评价体系。主要成果包括：  （1）基于Nesterov梯度优化和放缩不变性的对抗攻击方法  发现了梯度累积的重要性和图像对模型分类的伸缩不变性，首次将对抗样本生成过程与模型标准训练过程进行类比分析并被业界广泛接受，分别提出基于Nesterov梯度优化和图像伸缩不变性的数据增强对抗攻击方法，利用当前样本点的梯度向前优化下一步生成的梯度并对图像进行放缩生成多个副本，增强了黑盒攻击的迁移攻击成功率。被中国科学院院士谭铁牛及多位IEEE会士等正面引用。  （2）基于方差缩减的对抗攻击迁移性方法  发现了生成对抗样本的过程中，生成的梯度存在方差较大的现象，首次提出通过在对抗样本点周围采样的方式，利用采样点的平均梯度生成对抗扰动，从而生成方向更加准确的梯度，提高了对抗迁移攻击的成功率。同时，该方法能够与现有的其他梯度优化和数据增强对抗攻击方法相结合。被多位IEEE会士等正面引用。  （3）基于单词显著性的文本分类深度学习模型对抗攻击方法  团队在自然语言处理领域有着深厚的积累，构造了实体关系自动抽取方法及语言处理技术平台，建立了基于词汇联想网络与实体关系词替换的隐蔽攻击方法。由于文本的离散性质和语义约束，与图像数据存在显著差异，其对抗样本的定义、对抗攻击与防御算法需要独立设计。作为最早研究文本对抗样本的少数团队，明确了文本对抗样本与图像对抗样本的区别与联系，完善了文本对抗样本的定义，提出了基于单词显著性的同义词替换攻击算法。该工作发表在ACL Oral，是文本对抗领域的早期、奠基性工作和模型鲁棒性的公认评估方法，被多位IEEE/ACM会士，以及斯隆研究奖得主Nicolas Papernot和Bo Li等引用。  依托项目成果构建了深度学习模型对抗攻防理论与技术体系，在CCF A/B类等权威会议和期刊发表论文40余篇。5篇代表作的谷歌学术总引用2200余次，得到中美欧澳新等多国院士、数十位ACM/IEEE会士、以及多位斯隆奖得主的正面引用和好评，为深度学习模型的安全可信性研究做出了创新性贡献。作为核心成员，参加与对抗机器学习相关的2项IEEE标准的制定（IEEE 3129-2023、IEEE 3168-2024），项目成果被集成到ARES、CANARY、OpenAttack、Adversarial-Attacks-PyTorch、MindArmour、TextAttack、PyABSA和Counterfit等多个著名开源框架，为阿里巴巴公司的商业治理核心业务模型保持指定准确率的前提下带来10%的召回率提升。项目成果产生较大的国际学术影响，第一完成人多次受邀在CCF ADL讲习班、CCF AI、CCDM等专委会议上作对抗机器学习的学术报告，并入选省网信优秀人才。 |
| 主要完成人  （完成单位） | 何琨（华中科技大学）、车万翔（哈尔滨工业大学）、林嘉栋（华中科技大学）、王晓森（华中科技大学）、  任抒怀（华中科技大学） |
| |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | 序号 | 论文（专著）名称/刊名/作者 | 年、卷、页码 | 发表时间（年月日） | 通讯作者（含共同） | 第一作者（含共同） | 国内作者 | | 1 | “Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency”, Proceedings of the 57th Conference of the Association for Computational Linguistics, S. Ren (任抒怀) and Y. Deng and Kun He (何琨) and Wanxiang Che (车万翔) (CCF A类论文, CORE A\* 类论文) | 2019, pp. 1085-1097 | 2019-07 | 何琨 | 任抒怀 | 车万翔 | | 2 | “Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks”,8th International Conference on Learning Representations, Jiadong Lin (林嘉栋), C. Song , Kun He (何琨), L. Wang and J. E. Hopcroft (CORE A\* 类论文) | 2020 | 2020-04 | 何琨 | 林嘉栋 | 宋传标，王立威 | | 3 | “Enhancing the Transferability of Adversarial Attacks Through Variance Tuning”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Wang Xiaosen (王晓森) and Kun He (何琨)(CCF A类, CORE A\* 类论文) | 2021, pp. 1924-1933 | 2016-06 | 何琨 | 王晓森 | - | |  | “实体关系自动抽取”，中文信息学报，车万翔，刘挺，李生（CCF T1中文论文） | 2005，no.2, pp. 1-6 | 2005-03 | 车万翔 | 车万翔 | 李生，刘挺 | |  | “语言技术平台”， 中文信息学报，刘挺，车万翔，李正华（CCF T1中文论文） | 2011，vol. 25, no. 6, pp. 53-62 | 2011-11 | 车万翔 | 刘挺 | 李正华 | | |